AN EFFICIENT STORAGE MECHANISM FOR REPRESENTING TERM OCCURRENCE IN UNSTRUCTURED TEXT DOCUMENTS

ABSTRACT

A method and structure converts a document corpus containing an ordered plurality of documents into a compact representation in memory of occurrence data, where the representation is to be based on a dictionary previously developed for the document corpus and where each term in the dictionary has associated therewith a corresponding unique integer. The method includes developing a first vector for the entire document corpus, the first vector being a sequential listing of the unique integers such that each document in the document corpus is sequentially represented in the listing according to the occurrence in the document of the corresponding dictionary terms. A second vector is also developed for the entire document corpus and indicates the location of each of the document's representation in the first vector.